Hippocampal Cognitive Function Based on Deep Learning

Bijun Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1278004587@qq.com

Hongge Yao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 835092445@qq.com

***Abstract*—This research focuses on the study of agent behavior decision-making based on hippocampal cognitive functions, aiming to enhance the decision-making capabilities of agents in complex task environments by deeply exploring the crucial role of the hippocampus in learning, memory, and cognitive processes. By drawing inspiration from the biological structure and functional characteristics of the hippocampus, researchers are dedicated to designing and developing more intelligent and adaptive decision-making models to enhance agents' behavioral performance, problem-solving abilities, and adaptability to new situations. To achieve this goal, the research integrates advanced artificial intelligence technologies such as reinforcement learning and deep learning to simulate the complex functions of the hippocampus in memory encoding, storage, retrieval, and cognitive reasoning. This research not only contributes to advancing intelligent systems towards higher levels of intelligence and personalization but also plays a significant role in improving the interaction between intelligent agents and humans, providing intelligent services that better meet user needs. We found that the neural network trained in multi-task learning benefits from a loss term that promotes relevant and irrelevant representations. Therefore, the complementary coding we found in CA3 can provide extensive computational advantages for solving complex tasks. Furthermore, the study emphasizes the importance of further elucidating the functional mechanisms of the hippocampus, with the expectation of providing a more solid theoretical foundation for the optimization and refinement of agent decision-making models in the future.**

Keywords-Reinforcement Learning; Hippocampus; Memory Encoding

I. INTRODUCTION

The human brain is a general intelligence system consisting of hundreds of billions of neurons and millions of trillions of synaptic connections, endowed with the abilities of perception, learning, reasoning, and decision making. Cognition refers to the brain's perception, understanding, and memory of external stimuli, while decision-making involves the selection of actions based on cognitive information. Cognitive decision-making is the process of choosing the best course of action through thinking, analyzing, and evaluating information. It engages our capabilities of thought, perception, memory, and reasoning.

When making decisions, this paper may be influenced by cognitive biases, leading to irrational choices. In recent years, the intersection of cognitive neuroscience and artificial intelligence has become increasingly close, particularly in applying profound insights from neurobiology to decision-making systems in intelligent agents, where significant progress has been made. This trend is deeply inspired by the efficient decision-making abilities exhibited by humans and other advanced organisms in complex environments. These organisms can quickly make complex inferences from limited information and flexibly integrate new knowledge to optimize their behavior, a capability crucial for building more intelligent and adaptive intelligent agents.

The hippocampus, as the core region of the brain responsible for memory formation, storage, and retrieval, has unique cognitive functions that have become a key source of inspiration for designing decision-making models in intelligent agents. Researchers are dedicated to unraveling the complex structure and functions of the hippocampus, especially how it interacts with other brain regions (such as the Para hippocampal gyrus, parietal lobe, frontal lobe, and cerebral cortex) to support advanced cognitive tasks. By

simulating these intricate characteristics of the hippocampus, researchers aspire to develop advanced intelligent agent models that possess the capability to make precise and adaptable decisions within highly complex and ever-changing environments, mirroring the decision-making process exhibited by humans in their natural surroundings.

II. Related Works

A. Function and Morphology of the Hippocampus

Unlike the neocortex, the hippocampus and its adjacent dentate gyrus belong to the archicortex, featuring a three-layered cellular structure consisting of the molecular layer, the pyramidal cell layer, and the polymorphous cell layer. Based on its organizational characteristics, the hippocampus can be further divided into four regions: CA1, CA2, CA3, and CA4. CA1 and CA2 are located on the dorsal side of the hippocampus, while CA3 and CA4 are situated on the ventral side. The hippocampus, together with its nearby dentate gyrus, subiculum, parahippocampal gyrus, and cingulate gyrus, forms a structural and functional unity known as the hippocampal formation. The hippocampal formation has direct fiber connections with the septal area, entorhinal cortex, and the mamillary bodies of the hypothalamus through the fornix, fimbria of the hippocampus, and perforant path. The dentate gyrus of the hippocampal formation directly receives neural information from the amygdala, other limbic cortices, and the neocortex via the perforant path emanating from the entorhinal cortex. After receiving neural information from these brain structures, the dentate gyrus sends fibers to CA3 and CA4, from which axonal collaterals (Schaffer collateral fibers) of CA3 and CA4 neurons terminate in CA1 and CA2 of the hippocampus. Although the fornix primarily consists of efferent fibers from the hippocampal formation, it also contains cholinergic afferent fibers from the medial septal nucleus as well as serotonergic and noradrenergic fibers originating from the brainstem. The main efferent fibers of the hippocampal formation originate from the CA1 and CA2 regions, reaching the mamillary bodies of the hypothalamus, the anterior thalamic nuclei, and the lateral septal nucleus via the fornix. The efferent fibers from the CA1 and CA2 regions also terminate in the subiculum. Among these connections in the hippocampal formation, the majority of synapses use amino acid substances as neurotransmitters, primarily glutamate and GABA. Two noteworthy circuits are the classic Papaz's circuit and the trisynaptic circuit.

B. The trisynaptic memory circuit of the hippocampus

It was first reported by Lomo in 1966, who described a phenomenon he termed long-term potentiation (LTP) occurring in the trisynaptic circuit of the hippocampus. This discovery subsequently gained widespread attention due to its relevance to the brain mechanisms of memory. The trisynaptic circuit initiates within the entorhinal cortex, where neuronal axons coalesce to create the perforant pathway, ultimately terminating on the dendrites of granule cells located in the dentate gyrus. This constitutes the first synaptic link. Subsequently, the axons emanating from these granule cells in the dentate gyrus transform into mossy fibers, which establish synaptic connections with the dendrites of pyramidal cells residing in the CA3 area of the hippocampus, thereby forging the second synaptic junction.

The axons of the pyramidal cells in the CA3 region send collaterals to make the third synaptic connection with pyramidal cells in the CA1 region. From there, pyramidal cells in the CA1 region project back to the medial entorhinal cortex. This trisynaptic circuit, connecting the dentate gyrus, entorhinal cortex, and hippocampus, possesses unique functional properties and was initially considered evidence supporting the mechanism of long-term memory.

C. Small loop of hippocampal CA3

The hippocampus serves as the cornerstone of our ability to form episodic memories, enabling us to narrate personal experiences from our daily lives. The sensory information pertinent to memory storage travels through the entorhinal cortex (EC), functioning as the primary gateway or initiating point, serves as the cornerstone of the trisynaptic circuit. conduit between the

hippocampus and the neocortex. The anatomy and physiology of the hippocampus intertwine with fundamental attractor network theory, which encompasses two key findings: Tsodyks and Feigel's discovery the capacity to store sparse, uncorrelated patterns in abundance is complemented by Fontanari's insight, which suggests that dense, correlated patterns can coalesce into representations embodying shared characteristics. Both these forms of representation can harmoniously coexist and be retrieved within the same neural network, contingent upon a certain threshold being met. serving as the selector between them.

Neurons in layer II of the EC project to the CA3 region via two distinct pathways, as depicted in Figure 3. One pathway directly synapses with the distal dendrites of CA3 pyramidal cells via the perforant path (PP). The alternative route sees the PP axons branching off to the dentate gyrus (DG) before reaching CA3, where they form synapses with granule cells [1]. These granule cells, in turn, extend mossy fibers (MF) that establish synaptic connections with the closer, proximal dendrites of the pyramidal cells located in the CA3 region. Regarding the identical sensory data, two distinct representations emerge, each endowed with One is sparse and decorrelated, achieved through the mediation of mossy fibers (MFs), while the other is dense and correlated, facilitated by the perforant path (PP).

III. Algorithms Model

A. Description of the problem

The CA3 subregion of the hippocampus is recognized the hippocampus operates as an autoassociative network, encoding experiences into enduring memories. The raw data pertaining to these experiences stems both directly from the entorhinal cortex and indirectly, via the dentate gyrus which acts as a filter, performing sparsification and decorrelation. The computational goals pursued by these dual input routes can be rephrased as enhancing the efficiency and accuracy of memory encoding. have yet to be conclusively determined. Here, this project conceptualizes CA3 as a Hopfield-analogous network, proficient in accommodating both dense, correlated encodings and sparse, uncorrelated ones. As the number of memories accumulates, the dense encodings tend to coalesce around common features, while the sparse encodings maintain their individuality.

This project emulates the transformation of memory representations as they traverse the two pathways from the EC to CA3, and explore how these transformed encodings are subsequently stored and retrieved within CA3. Additionally, the hippocampus plays a pivotal role in recognizing similarities and patterns across disparate experiences, thereby enhancing cognitive processes. By modeling the hippocampal network, this work initially hypothesizes that MF (mossy fiber) encodings and PP (perforant path) encodings in CA3 can preserve distinctions between memories while enabling generalization among them [2]. Our goal is to delve into whether an auto associative network possesses the capability to preserve and recall memory encodings originating from both pathways, this paper conduct our investigation., enabling information representation at different scales that allows the network to both differentiate between instances and generalize across them. Through training an artificial neural network, this work ultimately demonstrates that these encoding types are suited to performing complementary tasks of instance recognition and concept classification. This approach enables a more intricate and nuanced comprehension. Hippocampus processes and integrates information, ultimately contributing to a deeper understanding of its functional role in memory storage and retrieval. contributes to memory formation and retrieval.

B. CA3-inspired Complementary Coding Model

Transition of Memory along the Hippocampal Pathway, from Image to Binary Autoencoder in EC, FNN Network from EC to CA3, visualizing Pathways from CA3 back to EC. The Hopfield-like Model in CA3.

The Hopfield neural network functions similarly to a memory storage device. When multiple sequences or images are input into this network, it stores this information in the form of

connection weights between neurons. Upon re-inputting the same or partially corrupted original sequence/image, the network is capable of restoring (recovering) the sequence/image. Figure 1 is a network model.

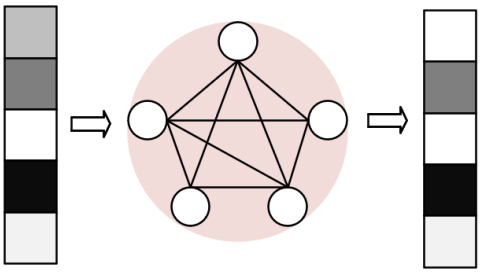

Figure 1. CA3 Hopfield-like model

The Hopfield-like network stores both sparse, uncorrelated encodings and dense, correlated encodings [3]. As more memories are stored, the former tend to remain distinct, while the latter merge along shared features. During its dynamic evolution, the Hopfield network converges towards stable states, which are the attractors of the network [4]. The design of network weigh and the initial state determine which attractor the network ultimately converges to. At Figure 2. In an auto-associative network, the attractor basins of the former tend to remain independent, while those of the latter tend to merge.

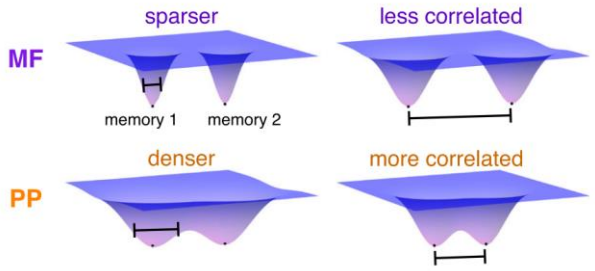

Figure 2. Basin of Attraction

Pattern storage: The Hopfield-like model of CA3 serves as a mechanism for pattern storage, where it retains the information by encoding the linear combination of patterns originating from the medial entorhinal cortex (MF) and the perforant path (PP).

$$q_{\mu\nu i} = (1-\zeta)\cdot\left(x_{\mu\nu i}^{MF} - a_{MF}\right) + \zeta\cdot\left(x_{\mu\nu i}^{PP} - a_{PP}\right) \quad (1)$$

The PP pattern $\zeta = 0.1$ stands out in terms of its comparative intensity. A defining aspect of Hopfield networks with binary neural states.0 and 1 is the subtraction of a density value from each pattern. The PP inputs have a notably weaker intensity compared to others, stemming from their more remote positioning (PP distal synapses) and the fact that they are empirically weaker than MF synapses (which are located on proximal dendrites).

These inputs undergo linear summation and are Architecture that is defined by its interconnectivity pattern, with i and j serving as indices for post-synaptic and pre-synaptic neurons, respectively.

$$W_{ij} \sim \sum_{\mu\nu}\left(0.9x_{\mu\nu i}^{MF} + 0.1x_{\mu\nu i}^{PP}\right)\left(0.9x_{\mu\nu j}^{MF} + 0.1x_{\mu\nu j}^{PP}\right) \quad (2)$$

The process of pattern retrieval involves generating a cue by randomly altering the activation state of 0.01 of the neurons in the target pattern, a quantity that is termed cue inaccuracy [5]. Throughout the retrieval phase, neurons undergo asynchronous updates in iterative cycles, where each neuron is updated once per cycle in a random sequence. At any specific instant in time, denoted as t, the cumulative synaptic input represents the aggregated electrical signals received by a neuron from its presynaptic counterparts. stored within a Hopfield-like network.

$$g_i(t) = \sum_j W_{ij} S_j(t) + h_i(t) \quad (3)$$

C. Model Loss function

The conversion of memory through the hippocampal pathway involves an image being encoded into a binary autoencoder form within the EC. In Figure 3, this project constructed and trained a comprehensive fully connected linear autoencoder architecture, comprising three strategically sized hidden layers (128, 1024, 128), each tailored to facilitate efficient information encoding and decoding.

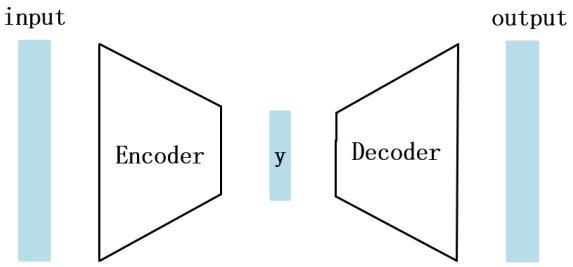

Figure 3. AE Network Architecture

To ensure swift and stable learning dynamics, this work incorporated batch normalization into each layer. Mitigating internal covariate shift and accelerating the training process. For nonlinearity, the ReLU function was applied to the first and third hidden layers, while the Sigmoid function was reserved for the output layer. Critically, the activations within the intermediate hidden layer underwent binarization through the Heaviside step function, with gradient flow maintained during backpropagation via the straight-through estimator. The overall optimization was guided by a specified loss function.

$$\mathcal{L}=\sum_{batch}^{\mu\nu\in}\left\|\hat{\imath}_{\mu\nu}-i_{\mu\nu}\right\|^2+\lambda\sum_{batch}^{\mu\nu\in}KL\left(\frac{1}{N_{EG}}\sum_i x_{\mu\nu i}^{EC}\middle\|a_{EC}\right) \quad (4)$$

I represent the original images, comprising pixel intensities spanning a spectrum from 0 to 1, and its reconstructed counterpart. x denotes the binary activations within the intermediate hidden layer, facilitating a sparse representation. And with an expected density of a=0.1, this paper employ sparsification with an intensity that evaluates the Kullback-Leibler (KL) divergence—a metric comparing the density of the hidden layer activations to the desired target density. Through this process, this project obtains the desired λ value of 10, which effectively achieves the expected sparsity level.

The central characteristic of the CA3 model lies in its activity threshold, which dictates whether the network retrieves example-based encodings or concept-based encodings. The project postulates the theta oscillations in the CA3 region, as a fundamental neural rhythm, not only embody pivotal threshold but also dynamically. Orchestrate fine-tuning of memory retrieval, enabling the brain to access and retrieve information from a broader or narrower range of memories, depending on the specific context and cognitive demands. This process is intricately intertwined with synaptic plasticity and network connectivity, facilitating the adaptive adjustment of memory representations to better serve the organism's current needs and goals. The work introduces a plug-and-play loss function. That endows artificial neural networks with the comprehensive ability to represent both complex and diverse data patterns. pattern-separated (PP) and pattern-completed (MF) classes. Compared to networks with solely rely on a single representation type, these networks, by virtue of their ability to integrate diverse information through multiple representation types exhibit superior performance in multitask learning.

The DeCorr loss function addresses the issues of oversmoothing and excessive feature correlation by reducing the correlation between features. Consequently, the paper applies the DeCorr loss function to decorrelate encodings in the final hidden layer, mimicking the MF (sparse and decorrelated) mode observed in CA3. The exclusion of the encoding loss function ensures that the encoded representations preserve the intrinsic image correlations and patterns. DeCorr simulates the MF pathway by considering the baseline condition where there is no loss function applied to hidden layer activations, thus preserving the natural correlations between similar images and mimicking the PP pathway.

$$\mathcal{L}_{DeCorr}\approx\frac{1}{2}\sum_{\alpha,\beta\in batch}Pearson\left(s_\alpha,s_\beta\right)^2 \quad (5)$$

It has been observed that different encoding properties are suited for distinct tasks. The baseline network excels in conceptual learning, whereas the DeCorr network typically performs better in exemplar learning but struggles with conceptual learning. To address this, the project applies the HalfCorr loss function, which decorrelates encodings only in the latter half of the final hidden layer. The introduction of the HalfCorr loss function diversifies the hidden layer representations, incorporating both correlated and uncorrelated components. As a result, HalfCorr networks are better equipped to learn tasks that

involve distinguishing between similar inputs and generalization.

HalfCorr networks demonstrate high performance in both tasks. Drawing parallels to the CA3 model, the paper find that exemplars are the decorrelated MF pathway is biased towards. Encoding information in a way that enhances discrimination and minimizes overlap among different elements. while concepts are preferentially encoded correlated PP pathway.

$$\mathcal{L}_{HalfCorr} \approx \frac{1}{2}\sum_{\alpha,\beta\in batch} Pearson\left(s_{\alpha}^{half}, s_{\beta}^{half}\right)^2 \quad (6)$$

s_{α}^{half} representing the latter half of the neurons in the final hidden layer. The DeCorr network excels in exemplar learning but suffers from inferior performance in conceptual learning, a trade-off that does not affect the HalfCorr network. The HalfCorr network displays high performance in both tasks, demonstrating an ability to prioritize the use of each type of encoding for tasks it is better suited for. The work comprehensively quantifies the impact of individual neurons on various tasks by precisely measuring the decrement in task accuracy that ensues upon their silencing. This approach offers a nuanced understanding of how each neuron contributes to the overall performance. Furthermore, DeCorr, an innovative technique, empowers us to delicately modulate the encoding correlations within artificial neural networks, thereby amplifying the salience of input features that are crucial for accurate predictions.

By strategically aligning the computational requirements of diverse tasks with the optimal encoding scales tailored for each, DeCorr facilitates a more efficient resolution of these tasks. This alignment ensures that the network's resources are allocated effectively, enhancing both speed and accuracy. Notably, correlated neurons within these networks exhibit a pronounced influence on conceptual learning, facilitating the extraction of abstract representations that generalize across examples. Conversely, decorrelated neurons play a pivotal role in exemplar learning, capturing specific details that distinguish individual instances within a category. By leveraging the complementary strengths of correlated and decorrelated neurons, DeCorr promotes a balanced and flexible learning strategy that is well-suited to tackle a wide range of complex tasks. This approach not only advances our theoretical understanding of neural network behavior but also has practical implications for designing more efficient and robust machine learning systems. The work proposes a distinct paradigm where loss functions are applied to distinct neurons

Fosteringthe principle of heterogeneity within a layer can be enriched by tailoring the degree of decorrelation for individual components or clusters within the HalfCorr network [6].

IV. Experiments

A. Experimental Environment

In the experiments, the performance of the cognitive algorithm inspired by hippocampal memory designed in this paper is evaluated, and its properties are analyzed [7].

Firstly, the sample efficiency project undertakes a comparative assessment to gauge how the novel algorithm fares against previous conventional neural networks. All experiments are implemented on an RTX3060 GPU with 16GB of VRAM and a CPU running at 14.4 GHz, utilizing Pytorch and NVIDIA CUDA.

B. Dataset

In our model utilizing the Fashion-MNIST dataset, the sensory input encoded as memory consists of Fashion-MNIST images. The memory comprises 256 images from each of the categories of sneakers, trousers, and coats. These memories, which are Fashion-MNIST images, serve as exemplars representing individual concepts.

C. Train the network

To evaluate the performance of a cognitive algorithm inspired by hippocampal memory, the project compared it with traditional classification and recognition algorithms using the MNIST dataset. The paper normalized images, randomly assigned set numbers, and trained a multi-layer perceptron to either classify digits or identify sets. The network was trained on a subset of images

and evaluated on a test set for digit classification and on corrupted images from the training set for set identification. Classification requires clustering images based on common features, akin to concept learning in our CA3 model, while distinguishing differences among similar images necessitates example learning, similar to our CA3 model [8].The project used stochastic gradient descent with a batch size of 50 and a learning rate of 1e−4.

Comparative Experiment:

D. Some Common Mistakes

- Traditional Classification Algorithms: The work utilizes the same dataset to train various traditional classification algorithms, including but not limited to Support Vector Machines, Decision Trees, and Random Forests, and subsequently evaluate and compare their performance on the test set.
- Hippocampal Memory Mechanism Simulation: In addition to directly training hippocampal-inspired cognitive model for classification, one can also attempt to introduce mechanisms into the model that simulate characteristics of the hippocampus, such as employing specific loss functions or regularization terms to encourage the network to learn sparse, uncorrelated representations (analogous to mossy fiber (MF) coding) or dense, correlated representations (analogous to perforant path (PP) coding) [9].

Dots indicate means, bars show SD of networks. The DeCov loss, developed to reduce overfitting, aids numerical convergence. DeCorr decorrelates input pairs for all neurons in a layer, while DeCov decorrelates neuron pairs across all inputs. At Figure 4 and Figure 5, as a generalization-boosting regularizer, DeCov enhances digit classification but not significantly set identification, contrasting DeCorr's effect. DeCorr impairs concept learning but boosts instance learning. The paper trained an MLP for concurrent digit classification & set identification. Compared to baselines, DeCorr networks often excel in instance learning but underperform in concept learning. The project train until >99.9% accuracy on the training set.

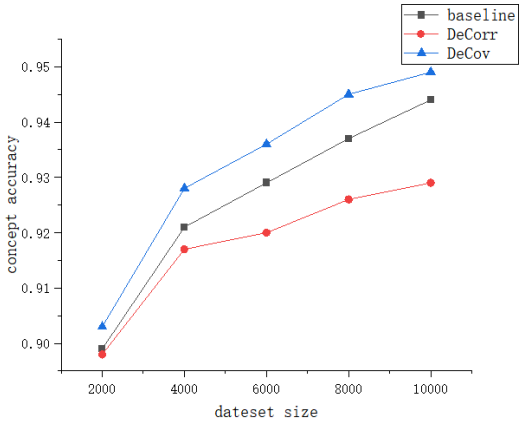

Figure 4. Comparison Chart of DeCov under MF Modes

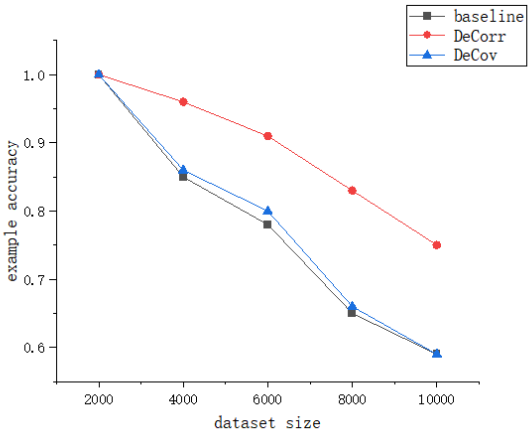

Figure 5. Comparison Chart of DeCov under PP Modes

Using the decrease in task accuracy after neuron silencing as an indicator of its impact, the work found that correlated neurons have a greater influence on concept learning, while decorrelated neurons impact instance learning more significantly. The average drop in accuracy across each neuron in the network reveals their respective contributions to both learning modalities. For all results, p-values were calculated using an unpaired, two-tailed t-test.

As can be seen from Figure 6 and Figure 7, Correlated neurons (orange bars) exhibit a stronger influence on concept learning can reach 0.00077613, whereas decorrelated neurons (purple bars) have a more pronounced effect on instance learning) can reach 0.0037056 [10]. For all results,

p-values were calculated using an unpaired, two-tailed t-test.

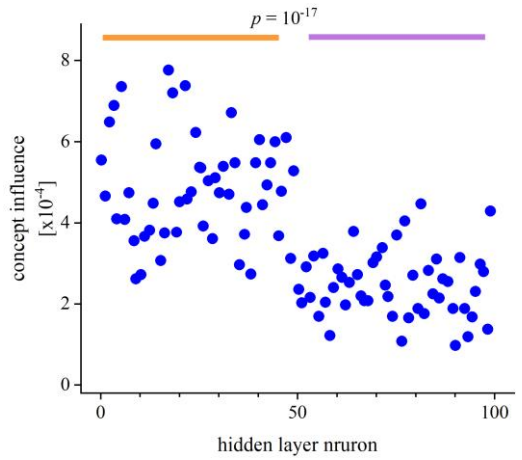

Figure 6. The impact of neurons on concept learning

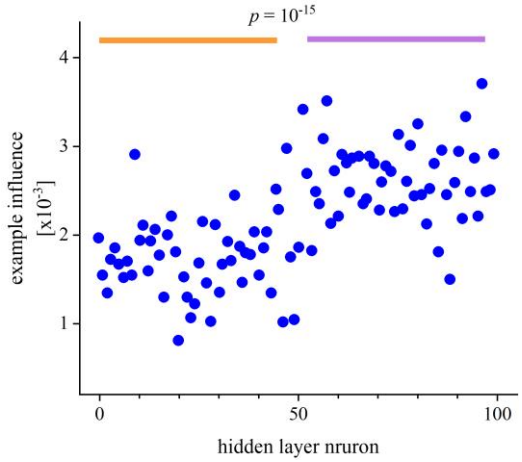

Figure 7. The impact of neurons on exemplar learning

V. CONCLUSIONS

In this paper, these novel agent models will not merely enhance performance in specific tasks but also propel intelligent systems towards greater intelligence, adaptability, and personalization. They will excel at comprehending and adapting to ever-evolving environments, offering more tailored and user-centric intelligent services. Furthermore, this interdisciplinary integration will pave new avenues for improving the interaction between intelligent systems and humans, fostering a harmonious coexistence between man and machine.

However, it is crucial to acknowledge that despite remarkable advancements, the precise functions of the hippocampus and its intricate relationship with overall cognitive processes remain a complex and incompletely unraveled domain. As such, future research endeavors will continue to delve deeper into the working mechanisms of the hippocampus, aiming to refine and optimize agent decision-making models, thereby propelling artificial intelligence technology to even greater heights.

REFERENCES

[1] Borzello, M. Assessments of dentate gyrus function: discoveries and debates. Nat. Rev. Neurosci. 24,502–517(2023).

[2] Asutay, E. Affective calculus: the construction of affect through information integration over time. Emotion 21,159–174 (2019).

[3] Herweg, N. A., Solomon, E. A. & Kahana, M. J. Theta oscillations in human memory. Trends in Cognitive Sciences 24,208–227 (2020).

[4] L. Kang and T. Toyoizumi. Hopfield-like network with complementary encodings of memories. Phys. Rev. E, 108(5):054410, 2023.

[5] Zheng, J. Multiplexing of theta and alpha rhythms in the amygdala-hippocampal circuit supports pattern separation of emotional information. Neuron 102,887–898 (2019).

[6] Barry, D. N. & Love, B. C. A neural network account of memory replay and knowledge consolidation. Cereb. Cortex. 33, 83–95(2022).

[7] Xiao, H., Rasul, K., & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms arXiv 1708.07747 (2017).

[8] Qasim, S. E., Fried, I. & Jacobs, J. Phaseprecessioninthe human hippocampus and entorhinal cortex. Cell 184,3242–3255 (2021).

[9] Vertes, E., and Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. Adv. Neural Inf. Process. Syst. 32, 13714–13724.

[10] Sun, C., Yang, W., Martin, J., and Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. Nat. Neurosci. 23,651–663.